Elements of Asymptotic Theory

James L. Powell
Department of Economics
University of California, Berkeley

Objectives of Asymptotic Theory

While exact results are available for, say, the distribution of the classical least squares estimator for the normal linear regression model, and for other leading special combinations of distributions and statistics, generally distributional results are unavailable when statistics are nonlinear in the underlying data and/or when the joint distribution of those data is not parametrically specified. Fortunately, most statistics of interest can be shown to be smooth functions of sample averages of some observable random variables, in which case, under fairly general conditions, their distributions can be approximated by a normal distribution, usually with a covariance matrix that shrinks to zero as the number of random variables in the averages increases. Thus, as the sample size increases, the gap between the true distribution of the statistic and its approximating distribution asymptotes to zero, motivating the term *asymptotic theory*.

There are two main steps in the approximation of the distribution of a statistic by a normal distribution. The first step is the demonstration that a sample average converges (in an appropriate sense) to its expectation, and that a standardized version of the statistic has distribution function that converges to a standard normal. Such results are called *limit theorems* – the former (approximation of a sample average by its expectations) is called a *law of large numbers* (abbreviated "LLN"), and the second (approximation of the distribution of a standardized sample average by a standard normal distribution) is known as a *central limit theorem* (abbreviated "CLT"). There are a number of LLNs and CLTs in the statistics literature, which impose different combinations of restrictions on the number of moments of the variables being averaged and the dependence between them; some basic versions are given below.

The second step in deriving a normal approximation of the distribution of a smooth function of sample averages is demonstration that this smooth function is approximately a linear function of sample averages (which is itself a sample average). Useful results for this step, which use the usual calculus approximation of smooth (differentiable) functions by linear functions, are grouped under the heading *Slutsky theorems*, after the author of one of the particularly useful results. So, as long as the conditions of the limit theorems and Slutsky theorems apply, smooth functions of sample averages are approximately sample averages

themselves, and thus their distribution functions are approximately normal. The whole apparatus of asymptotic theory is made up of the regularity conditions for applicability of the limit and Slutsky theorems, along with the rules for mathematical manipulation of the approximating objects.

In order to discuss approximation of a statistic by a constant, or its distribution by a normal distribution, we first have to extend the usual definitions of limits of deterministic sequences to accomodate sequences of random variables (yielding the concept of "convergence in probability") or their distribution functions ("convergence in distribution").

Convergence in Distribution and Probability

For most applications of asymptotic theory to statistical problems, the objects of interest are sequences of random vectors (or matrices), say $\mathbf{X}_N$ or $\mathbf{Y}_N$, which are typically statistics indexed by the sample size N; the object is to find simple approximations for the distribution functions of $\mathbf{X}_N$ or $\mathbf{Y}_N$ which can be made arbitrarily close to their true distribution functions for N large enough. Writing the cumulative distribution function of $\mathbf{X}_N$ as

$$F_N(\mathbf{x}) \equiv \Pr\{\mathbf{X}_N \leq \mathbf{x}\}$$

(where the argument $\mathbf{x}$ has the same dimension as $\mathbf{X}_N$ and the inequality holds only if it is true for each component), we define *convergence in distribution* (sometimes called *convergence in law*) in terms of convergence of the function $F_N(\mathbf{x})$ to some fixed distribution function $F(\mathbf{x})$ for any point $\mathbf{x}$ where the F is continuous. That is, the random vector $\mathbf{X}_N$ *converges in distribution to* $\mathbf{X}$, denoted

$$\mathbf{X}_N \xrightarrow{d} \mathbf{X} \qquad \text{as} \qquad \mathrm{N} \to \infty,$$

if

$$\lim_{N\to\infty} F_N(\mathbf{x}) = F(\mathbf{x})$$

at all points $\mathbf{x}$ where $F(\mathbf{x})$ is continuous.

A few comments on this definition and its implications are warranted:

1. The "limiting" random vector $\mathbf{X}$ in this definition is really just a placeholder for the distribution function $F(\mathbf{x})$; there needs be no "real" random vector $\mathbf{X}$ which has the limiting distribution function. In fact, it is often customary to replace $\mathbf{X}$ with the limiting distribution function $F(\mathbf{x})$ in the

definition, i.e., to write

$$\mathbf{X}_N \xrightarrow{d} F(\mathbf{x}) \qquad \text{as} \qquad \text{N} \to \infty;$$

if, say, $F(\mathbf{x})$ is a multivariate standard normal, then we would write

$$\mathbf{X}_N \xrightarrow{d} N(\mathbf{0}, \mathbf{I}),$$

where the condition "$N \to \infty$" is usually not stated, but implied.

2. If $\mathbf{X}_N \xrightarrow{d} F(\mathbf{x})$, then we can use the c.d.f. $F(\mathbf{x})$ to approximate probabilies for $\mathbf{X}_N$ if N is large, i.e.,

$$\Pr\{\mathbf{X}_N \in A\} \stackrel{A}{=} \int_A dF(\mathbf{x}),$$

where the symbol "$\stackrel{A}{=}$" means that the difference in the left- and right-hand sides goes to zero as N increases.

3. We would like to construct statistics so that their limiting distributions are of convenient (tabulated) form, e.g., normal or chi-squared distribution. This often involves *standardizing* the statistics by subtracting their expectations and dividing by their standard deviations (or the matrix equivalent).

4. If $\lim_{N\to\infty} F_N(\mathbf{x})$ is discontinuous, it may not be a distribution function, which is why the qualification "at all points where $F(\mathbf{x})$ is continuous" has to be appended to the definition of convergence in distribution. To make this more concrete, suppose

$$X_N \sim N\left(0, \frac{1}{N}\right),$$

so that

$$\begin{aligned} F_N(x) &= \Phi\left(\sqrt{N}x\right) \\ &\to \begin{cases} 0 \text{ if } x < 0 \\ \frac{1}{2} \text{ if } x = 0 \\ 1 \text{ if } x > 0 \end{cases}. \end{aligned}$$

Obviously here $\lim_{N\to\infty} F_N(x)$ is not right-continuous, so we would never use it to approximate probabilities for X_N (since it isn't a c.d.f.). Instead, a reasonable approximating function for $F_N(x)$ would be

$$F(x) \equiv 1\{x \geq 0\}.$$

where "$1\{A\}$" is the indicator function of the statement A, i.e., it equals one if A is true and is zero otherwise. The function $F(x)$ is the c.d.f. of a degenerate random variable X which equals zero with probability one, which is a pretty good approximation for $F_N(x)$ when N is large for this problem.

The last example – convergence of the c.d.f. $F_N(\mathbf{x})$ to the c.d.f. of a degenerate ("nonrandom") random variable – is a special case of the other key convergence concept for asymptotic theory, *convergence in probability.* Following Ruud's text, we define convergence in probability of the sequence of random vectors $\mathbf{X}_N$ to the (nonrandom) vector $\mathbf{x}$, denoted

$$\operatorname*{plim}_{N\to\infty} \mathbf{X}_N = \mathbf{x}$$

or

$$\mathbf{X}_N \xrightarrow{p} \mathbf{x} \qquad \text{as} \qquad N \to \infty$$

if

$$\mathbf{X}_N \xrightarrow{d} x,$$

i.e.,

$$F_N(\mathbf{c}) \equiv \Pr\{\mathbf{X}_N \leq \mathbf{c}\} \to \mathbf{1}\{\mathbf{x} \leq \mathbf{c}\}$$

whenever all components of $\mathbf{x}$ are different from those of $\mathbf{c}$. (The symbol "*plim*" stands for the term "*probabilty limit*".) We can generalize this definition of convergence in probability of $\mathbf{X}_N$ to a constant $\mathbf{x}$ to permit $\mathbf{X}_N$ to converge in probability to a nondegenerate random vector $\mathbf{X}$ by specifying that

$$\mathbf{X}_N \xrightarrow{p} \mathbf{X}$$

means that

$$\mathbf{X}_N - \mathbf{X} \xrightarrow{d} \mathbf{0}.$$

An equivalent definition of convergence in probability, which is the definition usually seen in textbooks, is that

$$\mathbf{X}_N \xrightarrow{p} \mathbf{x}$$

if, for any number $\varepsilon > 0$,

$$\Pr\{||\mathbf{X}_N - \mathbf{x}|| > \varepsilon\} \to 0$$

as $N \to \infty$. Thus, $\mathbf{X}_N$ converges in probability to $\mathbf{x}$ if all the probability mass for $\mathbf{X}_N$ eventually ends up in an ε-neighborhood of $\mathbf{x}$, no matter how small ε is. Showing the equivalence of these two definitions of probability limit is a routine exercise in probability theory and real analysis.

A stronger form of convergence of $\mathbf{X}_N$ to a nonrandom vector $\mathbf{x}$, which can often be used to show convergence in probability, is *quadratic mean convergence.* We say

$$\mathbf{X}_n \stackrel{qm}{\to} \mathbf{x}$$

as $N \to \infty$ if

$$E[||\mathbf{X}_N - \mathbf{x}||^2] \to 0.$$

If this condition can be shown, then it follows that $\mathbf{X}_N$ converges in probability to $\mathbf{x}$ because of

Markov's Inequality: If Z is a nonnegative (scalar) random variable, i.e., $\Pr\{Z \leq 0\} = 1$, then

$$\Pr\{Z > K\} \leq \frac{E(Z)}{K}$$

for any $K > 0$.

The proof of this inequality is worth remembering; it is based upon a decomposition of $E(Z)$ (which may be infinite) as

$$\begin{aligned} E(Z) &\equiv \int_0^\infty z dF_Z(z) \\ &= \int_0^K z dF_Z(z) + \int_K^\infty (z - K) dF_Z(z) + \int_K^\infty K dF_Z(z). \end{aligned}$$

Since the first two terms in this sum are nonnegative, it follows that

$$E(Z) \geq \int_K^\infty K dF_Z(z) = K \cdot \Pr\{Z > K\},$$

from which Markov's inequality immediately follows. A better-known variant of this inequality, *Tchebysheff's Inequality*, takes $Z = (X - E(X))^2$ for a scalar random variable X (with finite second moment) and $K = \varepsilon^2$ for any $\varepsilon > 0$, from which it follows that

$$\Pr\{|X - E(X)| > \varepsilon\} = \Pr\{(X - E(X))^2 > \varepsilon^2\} \leq \frac{Var(X)}{\varepsilon^2},$$

since $Var(X) = E(Z)$ for this example. A similar consequence of Markov's inequality is that *convergence in quadratic mean implies convergence in probability:* if

$$E[||\mathbf{X}_N - \mathbf{x}||^2] \to 0,$$

then

$$\Pr\{||\mathbf{X}_N - \mathbf{x}|| > \varepsilon\} \leq \frac{E[||\mathbf{X}_N - \mathbf{x}||^2]}{\varepsilon^2} \to 0$$

for any $\varepsilon > 0$.

There is another, stronger version of convergence of a random variable to a limit, known as *almost sure convergence* (or *convergence almost everywhere* or *convergence with probability one*), which is denoted

$$\mathbf{X}_N \overset{a.s.}{\to} \mathbf{x}.$$

Rather than give the detailed definition of this form of convergence – which is best left to a more advanced course – it is worth pointing out that it is indeed stronger than convergence in probability (which is sometimes called *weak convergence* for that reason), but that, for the usual purposes of asymptotic theory, the weaker concept is sufficient.

Limit Theorems

With the concepts of convergence in probability and distribution, we can more precisely state how a sample average is approximately equal to its expectation, or how its standardized version is approximately normal. The first result, on convergence in probability, is

Weak Law of Large Numbers (WLLN): If $\bar{X}_N \equiv \frac{1}{N}\sum_{i=1}^N X_i$ is a sample average of scalar, i.i.d. random variables $\{X_i\}_{i=1}^N$ which have $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ finite, then

$$\bar{X}_N \overset{p}{\to} \mu = E(X_i).$$

Proof: By the usual calculations for means and variances of sums of i.i.d. random variables, $E(\bar{X}_N) = \mu$, $Var(\bar{X}_N) = \frac{\sigma^2}{N} = E(\bar{X}_N - \mu)^2$, which tends to zero as $N \to \infty$. By definition, $\bar{X}_N$ tends to μ in quadratic mean, and thus in probability.

Inspection of the proof of the WLLN suggests a more general result for averages of random variables that are not i.i.d. – namely, that

$$\bar{X}_N - E(\bar{X}_N) \overset{p}{\to} 0$$

if

$$\frac{1}{N^2}\sum_{i=1}^N \sum_{j=1}^N Cov(X_i, X_j) \to 0.$$

This general condition can be specialized to yield other LLNs for heterogeneous (non-constant variances) or dependent (nonzero covariances) data. Other laws of large numbers exist which relax conditions on the existence of second moments, and which show almost sure convergence (and are known as "strong laws of large numbers").

The next main result, on approximate normality of sample averages, again applies to i.i.d. data:

Central Limit Theorem (Lindeberg-Levy): If $\bar{X}_N \equiv \frac{1}{N}\sum_{i=1}^{N} X_i$ is a sample average of scalar, i.i.d. random variables $\{X_i\}_{i=1}^{N}$ which have $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ finite, then

$$\begin{aligned} Z_N &\equiv \frac{\bar{X}_N - E(\bar{X}_N)}{\sqrt{Var(\bar{X}_N)}} \\ &= \sqrt{N}\left(\frac{\bar{X}_N - \mu}{\sigma}\right) \\ &\xrightarrow{d} N(0,1). \end{aligned}$$

The proof of this result is a bit more involved, requiring manipulation of characteristic functions, and will not be presented here. The result of this CLT is often rewritten as

$$\sqrt{N}(\bar{X}_N - \mu) \xrightarrow{d} N(0, \sigma^2),$$

which points out that $\bar{X}_N$ converges to its mean μ at exactly the rate $\sqrt{N}$ increases, so that the product eventually "balances out" to yield a random variable with a normal distribution. Another way to rewrite the consequence of the CLT is

$$\bar{X}_N \overset{A}{\sim} N\left(\mu, \frac{\sigma^2}{N}\right),$$

where the symbol "$\overset{A}{\sim}$" means "is approximately distributed as" (or "is asymptotically distributed as"). It is very important here to make the distinction between the *limiting distribution* of the standardized mean Z_N, which is a standard normal, and the *asymptotic distribution* of the (non-standardized) sample mean $\bar{X}_N$, which is actually a sequence of approximating normal distributions with variances that shrink to zero at speed $1/N$. In short, a "limiting distribution" cannot depend upon N, which has passed to its (infinite) limit, while an "asymptotic distribution" can involve the sample size N.

Like the Weak Law of Large Numbers, the Lindeberg-Levy Central Limit Theorem admits a number of different generalizations (often named after their authors) which relax the assumption of i.i.d. data while

imposing various stronger restrictions on the existence of moments or the tail behavior of the true data distributions.

To this point, the WLLN and CLT apply only to scalar random variables $\{X_i\}$; to extend them to random vectors $\{\mathbf{X}_i\}$, we can use the intriguingly-named *Cramér-Wold device*, which is really a theorem that states that if the scalar random variable

$$Y_N \equiv \boldsymbol{\lambda}' \mathbf{X}_N$$

converges in distribution to

$$Y \equiv \boldsymbol{\lambda}' \mathbf{X}$$

for any fixed vector $\boldsymbol{\lambda}$ having the same dimension as $\mathbf{X}_N$ – that is, if

$$\boldsymbol{\lambda}' \mathbf{X}_N \xrightarrow{d} \boldsymbol{\lambda}' \mathbf{X}$$

for any $\boldsymbol{\lambda}$ – then the vector $\mathbf{X}_N$ converges in distribution to $\mathbf{X}$,

$$\mathbf{X}_N \xrightarrow{d} \mathbf{X}.$$

This result immediately yields a multivariate version of the Lindeberg-Levy CLT: a sample mean $\bar{\mathbf{X}}_N$ of i.i.d. random vectors $\{\mathbf{X}_i\}$ with $E(\mathbf{X}_i) = \boldsymbol{\mu}$ and $V(\mathbf{X}_i) = \boldsymbol{\Sigma}$ has

$$\sqrt{N}(\bar{\mathbf{X}}_N - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}) .$$

And, since convergence in probability is a special case of convergence in distribution, the Cramér-Wald device also immediately yields a multivariate version of the WLLN.

Slutsky Theorems

To extend the limit theorems for sample averages to statistics which are functions of sample averages, asymptotic theory uses smoothness properties of those functions (i.e., continuity and differentiability) to approximate those functions by polynomials, usually constant or linear functions. The simplest of these approximation results is the *continuity theorem*, which states that probability limits share an important property of ordinary limits – namely, *the probability limit of a continuous function is the value of that function evaluated at the probability limit:*

Continuity Theorem: If $\mathbf{X}_N \xrightarrow{p} \mathbf{x}_0$and the function $g(\mathbf{x})$ is continuous at $\mathbf{x} = \mathbf{x}_0$, then $g(\mathbf{X}_N) \xrightarrow{p} g(\mathbf{x}_0)$.

Proof: For any $\varepsilon > 0$, continuity of the function $g(\mathbf{x})$ at $\mathbf{x}_0$ implies that there is some $\delta > 0$ so that $||\mathbf{X}_N - \mathbf{x}_0|| \leq \delta$ implies that $||g(\mathbf{X}_N) - g(\mathbf{x}_0)|| \leq \varepsilon$. Thus

$$\Pr\{||g(\mathbf{X}_N) - g(\mathbf{x}_0)|| \leq \varepsilon\} \geq \Pr\{||\mathbf{X}_N - \mathbf{x}_0|| \leq \delta\}.$$

But the probability on the right-hand side converges to one because $\mathbf{X}_N \xrightarrow{p} \mathbf{x}_0$, so the probability on the left-hand side does, too.

With the continuity theorem, it should be clear that probability limits are more conveniently manipulated than, say, mathematical expectations, since, for example, if

$$\hat{\theta} \equiv \frac{\bar{Y}_N}{\bar{X}_N}$$

is an estimator of the ratio of the means of Y_i and X_i,

$$\theta_0 \equiv \frac{E(Y_i)}{E(X_i)} \equiv \frac{\mu_Y}{\mu_X},$$

then $\hat{\theta}_N \xrightarrow{p} \theta_0$ as long as $\mu_X \neq 0$, even though $E(\hat{\theta}_N) \neq \theta_0$ in general.

Another key approximation result is *Slutsky's Theorem,* which states that the sum (or product) of a random vector that converges in probability and another that converges in distribution itself converges in distribution:

Slutsky's Theorem: If the random vectors $\mathbf{X}_N$ and $\mathbf{Y}_N$ have the same length, and if $\mathbf{X}_N \xrightarrow{p} \mathbf{x}_0$ and $\mathbf{Y}_N \xrightarrow{d} \mathbf{Y}$, then $\mathbf{X}_N + \mathbf{Y}_N \xrightarrow{d} \mathbf{x}_0 + \mathbf{Y}$ and $\mathbf{X}_N'\mathbf{Y}_N \xrightarrow{d} \mathbf{x}_0'\mathbf{Y}$.

The typical application of Slutsky's theorem is to estimators that can be represented in terms of a linear approximation involving some statistics that converge either in probability or distribution. To be more concrete, suppose the (scalar) estimator $\hat{\theta}_N$ can be decomposed as

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \mathbf{X}_N'\mathbf{Y}_N + Z_N,$$

where

$$\begin{aligned}
&\mathbf{X}_N \xrightarrow{p} \mathbf{x}_0, \\
&\mathbf{Y}_N \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}), \qquad \text{and} \\
&Z_N \xrightarrow{p} 0.
\end{aligned}$$

This decomposition is often obtained by a first-order Taylor's series expansion of $\hat{\theta}_N$, and the components $\mathbf{X}_N$, $\mathbf{Y}_N$, and Z_N are sample averages to which a LLN or CLT can be applied; an example is the so-called "delta method" discussed below. Under these conditions, Slutsky's theorem implies that

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{x}_0' \mathbf{\Sigma} \mathbf{x}_0).$$

Both the continuity theorem and Slutsky's theorem are special cases of the *continuous mapping theorem*, which says that convergence in distribution, like convergence in probability, interchanges with continuous functions, as long as they are continuous everywhere:

Continuous Mapping Theorem: If $\mathbf{X}_N \xrightarrow{d} \mathbf{X}$ and the function $g(\mathbf{x})$ is continous for all $\mathbf{x}$, then $g(\mathbf{X}_N) \xrightarrow{d} g(\mathbf{X})$.

The continuity requirement for $g(\mathbf{x})$ can be relaxed to hold only on the support of $\mathbf{X}$ if needed. The continuous mapping theorem can be used to get approximations for the distributions of test statistics based upon asymptotically-normal estimators; for example, if

$$\mathbf{Z}_N \equiv \sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}),$$

then

$$\begin{aligned} T &\equiv \mathbf{Z}_N' \mathbf{Z}_N = N(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)'(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \\ &\xrightarrow{d} \chi^2_p, \end{aligned}$$

where $p \equiv \dim\{\hat{\boldsymbol{\theta}}_N\}$.

The final approximation result discussed here, which is an application of Slutsky's theorem, is the so-called *"delta method"*, which states that continuously-differentiable functions of asymptotically-normal statistics are themselves asymptotically normal:

The "Delta Method": If $\hat{\boldsymbol{\theta}}_N$ is a random vector which is asymptotically normal, with

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma})$$

for some $\boldsymbol{\theta}_0$, and if $g(\boldsymbol{\theta})$ is continuously differentiable at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with Jacobian matrix

$$\mathbf{G}_0 \equiv \frac{\partial g(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}$$

that has full row rank, then

$$\sqrt{N}(g(\hat{\boldsymbol{\theta}}_N) - g(\boldsymbol{\theta}_0)) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{G}_0\boldsymbol{\Sigma}\mathbf{G}_0'\right).$$

Proof: First suppose that $g(\boldsymbol{\theta})$ is scalar-valued. Since $\hat{\boldsymbol{\theta}}_N$ converges in probability to $\boldsymbol{\theta}_0$ (by Slutsky's Theorem), and since $\partial g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ is continuous at $\boldsymbol{\theta}_0$, the Mean Value Theorem of differential calculus implies that,

$$g(\hat{\boldsymbol{\theta}}_N) = g(\boldsymbol{\theta}_0) + \frac{\partial g(\theta_N^*)}{\partial\theta'} \cdot (\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0),$$

for some value of $\boldsymbol{\theta}_N^*$ between $\hat{\boldsymbol{\theta}}_N$ and $\boldsymbol{\theta}_0$ (with arbitrarily high probability). Since $\hat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$, it follows that $\boldsymbol{\theta}_N^* \xrightarrow{p} \boldsymbol{\theta}_0$, so

$$\frac{\partial g(\boldsymbol{\theta}_N^*)}{\partial\boldsymbol{\theta}'} \xrightarrow{p} \mathbf{G}_0 \equiv \frac{\partial g(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}.$$

Thus, by Slutsky's Theorem,

$$\begin{aligned}
\sqrt{N}(g(\hat{\boldsymbol{\theta}}_N) - g(\boldsymbol{\theta}_0)) &= \frac{\partial g(\boldsymbol{\theta}_N^*)}{\partial\boldsymbol{\theta}'}.\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \\
&\xrightarrow{d} \mathbf{G}_0 \cdot N(\mathbf{0}, \boldsymbol{\Sigma}) \\
&= N(\mathbf{0}, \mathbf{G}_0\boldsymbol{\Sigma}\mathbf{G}_0').
\end{aligned}$$

To extend this argument to the case where $g(\hat{\boldsymbol{\theta}}_N)$ is vector-valued, use the same argument to show that any (scalar) linear combination $\boldsymbol{\lambda}'g(\hat{\boldsymbol{\theta}})$ has the asymptotic distribution

$$\sqrt{N}(\boldsymbol{\lambda}'g(\hat{\boldsymbol{\theta}}_N) - \boldsymbol{\lambda}'g(\boldsymbol{\theta}_0)) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\lambda}'\mathbf{G}_0\boldsymbol{\Sigma}\mathbf{G}_0'\boldsymbol{\lambda}),$$

from which the result follows from the Cramér-Wald device.

Tests of Nonlinear Hypotheses

The "delta method" can be used to construct a large-sample test for a nonlinear hypothesis

$$H_0 : \mathbf{g}(\boldsymbol{\theta}_0) = \mathbf{0},$$

where $\mathbf{g}(\boldsymbol{\theta})$ is a differentiable function with

$$\mathbf{G}(\boldsymbol{\theta}) \equiv \frac{\partial\mathbf{g}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}$$

assumed to be continuous and have full row rank in a neighborhood of $\boldsymbol{\theta}_0$. (An example would be $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{G}_0\boldsymbol{\theta} - \boldsymbol{\gamma}_0$, which is sometimes called the "general linear hypothesis.")

Suppose $\hat{\boldsymbol{\theta}}_N$ is asymptotically normal,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

and supposie a consistent estimator of the asymptotic covariance matrix $\boldsymbol{\Sigma}$ is available,

$$\hat{\boldsymbol{\Sigma}} \xrightarrow{p} \boldsymbol{\Sigma}.$$

Since

$$\hat{\mathbf{G}} \equiv \frac{\partial \mathbf{g}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'} \xrightarrow{p} \mathbf{G}(\boldsymbol{\theta}_0) \equiv \mathbf{G}_0$$

by the continuity theorem, it follows from Slutsky's theorem and the continuous mapping theorem that the "Wald statistic"

$$\begin{aligned} W_N &\equiv N \left(\mathbf{g}(\hat{\boldsymbol{\theta}})\right)' \left[\hat{\mathbf{G}}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{G}}'\right]^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}}) \\ &\xrightarrow{d} \chi^2_r \end{aligned}$$

under the null hypothesis, where $r = \dim\{\mathbf{g}(\boldsymbol{\theta})\} = rank\{\mathbf{G}_0\}$. (This is sometimes called the "generalized Wald" statistic, since Wald proposed the statistic for the special case of the *maximum likelihood* estimator, which is asymptotically normal under general conditions.) The Wald test of the null hypothesis H_0 would reject when W_N exceeds the upper α critical value for a chi-squared random variable with r degrees of freedom.